Transposition of the Cocitation Method With a View to Classifying Web Pages

Camille Prime-Claverie, Michel Beigbeder, Thierry Lafouge

Transposition of the Co-citation method with a view to Classifying Web Pages

C. Prime-Claverie, M. Beigbeder
Laboratoire RIM - SIMMO
Ecole Nationale Supérieure des Mines de Saint-Etienne
158, cours Fauriel
42023 Saint-Etienne Cedex, FRANCE
prime,mbeig@emse.fr

T. Lafouge
Laboratoire URSIDOC
Université Claude Bernard Lyon 1
43, Bd du 11 novembre 1918
69622 Villeurbanne Cedex, FRANCE
lafouge@univ-lyon1.fr

Abstract

The Web is a huge source of information and one of the main problems facing users is finding documents which correspond to their requirements. Apart from the problem of thematic relevance, the documents retrieved by search engines do not always meet the users' expectations. The document may be too general, or conversely too specialized, or of a different type from what the user is looking for, etc. We think that adding metadata to pages can considerably improve the process of searching for information on the Web. This article presents a possible typology for Web sites and pages, as well as a method for propagating metadata values, based on the study of the Web graph and more specifically the method of co-citation in this graph.

1. Introduction

The role of the search engines available on the Web is to retrieve in the minimum amount of time the most relevant pages on a given subject. It uses traditional information retrieval system techniques particularly for the representation of documents and queries and for matching systems. The aim is twofold: to find relevant Web pages and then rank them according to relevance. The search engines come up against two major difficulties. The first, which is well known when searching for information using uncontrolled vocabulary as is the case with full text searching, concerns language-based issues such as synonymy and polysemy which lead either to noise or silence. The second is directly related to the heterogeneous nature of the Web. In contrast to databases working on homogeneous corpuses of documents, that is, sets of selected documents assembled by the same authority and sharing common properties, (collections of scientific articles, patents, etc.), the Web is a forum of free expression which develops in an anarchic manner. It is disorganized and contains totally heterogeneous resources as far as language, subject, level, type, target audience, etc. are concerned. In such a world, quite apart from the problem of thematic relevance, it is difficult to find resources which correspond to the need (Gravano, 2000). Take the example of a Spanish student and a Spanish researcher both of whom are looking for information in nuclear physics. The first will look at papers in Spanish at a fairly basic level, while the second will look for scientific articles probably written in English, and possibly also at calls for papers or other documents relating to his/her scientific activity.

Along with many, we think that the use of metadata could greatly improve information retrieval on the Web (Marchiori, 1998). We are aware that we cannot count on all resource authors to correctly assign the proper metadata values, because this requires time, skill and objectivity. To obtain a uniform and systematic description of resources, assigning metadata values should be the work of an information retrieval system done in the same way as documentation professionals carry out cataloguing and indexing tasks. Since the manual application of metadata values is very costly, and given the number of pages available on the Web together with their volatility, it is not possible to imagine that they be created by hand. It is therefore necessary to look for automated or semi-automated methods. The methods considered are based on the propagation of metadata. To start with, only part of the resources is selected to have metadata assigned to them manually. These metadata are then propagated to other resources.

The method of propagation that we propose is carried out after applying an agglomerative hierarchical clustering method on the corpus. With our approach, this method uses similarity based on the Web's hypertext structure with a metric which comes from scientometry.

Currently, two scientific communities are closely involved in the analysis of Web hyperlinks: scientometrics as well as computing specialists. The first group, whose aim is, among others, to structure the universe of knowledge from large volumes of information, and to study the equivalence between the concepts established through scientometric analysis and the Web graph (Ingwersen, 1998), (Björneborn and Ingwersen, 2001), (Egghe, 2000). In fact, as in the scientific publication network (Garfield, 1972), a hypertext link can lead to a reference and indicate an interesting relationship between the original page and the page to which it is pointing. One of the limits of the analogy is to consider all hypertext links as citation or reference links. In fact, one must take into account publicity links, and especially those links which are used to move about a Web site (internal navigation links). Computer specialists, on the other hand, use mathematical methods from graph theory in order to improve information retrieval on the Web. The ranking algorithms of Google (Brin and Page, 1998), and the discovery of common interest communities (Kumar et al., 1999) are among the best known applications.

This article will first take a look at metadata and their use on the Web (section 2), then propose a possible typology for Web sites and pages (section 3). Afterwards, it will put forward our method of propagation based on the study of the Web graph, or more precisely on co-citation method (section 4). Finally, it will present a propagation experiment for metadata values relative to the typology of the Web pages defined in section 3 (section 5).

2. Metadata and its use on the Web

2.1 Definition and origin

Metadata is literally data on data. More precisely, the metadata of a resource can be considered as a set of information that describes it and is useful for using it. Metadata trace their origins back to the first library or museum catalogues but the advent of the computer has greatly expanded their use. Before the first electronic documents, metadata was stored outside documents in files *(external metadata)*, but now, with digital publishing, metadata can be included directly in the documents, generally in the header. This is called *internal metadata*. The metadata can be intended for the end user or also for various intermediaries, and either comes directly from the authors or publishers, or from information professionals such as information specialists. There are different metadata types

- Descriptive metadata, representing the resource and its information content, (title, author, keywords, etc.)
- Administrative metadata, related to the management of resources (intellectual property, localisation, etc.)
- Technical metadata, useful for consulting resources (data concerning security, digitalisation, etc.)
- Conservation metadata used for archiving resources.

2.2 Current Use of Metadata on the Web

The different markup languages used on the Web, such as HTML and now XML, provide for inserting internal metadata in the document header. However, this feature is not often used, probably because its availability is not well known. Moreover, even when author metadata is used, it is often misused, either because honest authors are not familiar with them or lack objectivity, or because those who are familiar with metadata divert them from their initial aim to increase their own visibility. This is why the majority of search engines do not take them into account in their algorithms. Despite this, efforts towards standardization continue. One of the most suited to digital resources is the Dublin Core project (2003). It provides fifteen metadata elements to give a “bibliographic” description of electronic resources on the Web. They are independent of the application field, and are designed to describe documents as well as objects such as images, maps and music. The fifteen metadata elements concern:

- Content: Title, Description, Subject, Source, Coverage, Type, Relation;

- Intellectual Property: Creator, Contributor, Publisher, Rights;
- Version of the Resource: Date, Format, Identifier, Language.

The Dublin Core standard uses ten mandatory attributes to describe each element and the manner in which they should be used. For example, one of the attributes specifies whether the metadata element is optional or not, another specifies if it can have one or more occurrences etc. This project is above all a standard for describing metadata and takes little account of the way that values could be assigned to the element.

3. Proposal for a Possible Typology for Web Sites and Pages

Among the metadata proposed by the Dublin Core, we believe that it is the metadata elements: subject, type, coverage and language that would be useful to improve information search on the Web. Given the abundance and diversity of available resources, users *a priori* do not have specific information on the authors or publication dates of the resources that they are looking for. On the other hand, they know the subject and the type of the documents they need and the languages they are capable of reading.

While there has been much discussion on metadata standards within the computing community, discussions on assigning metadata values and the difficulties related to this task are virtually non-existent. Very few authors propose standards or control lists for evaluation. The subject field of documents can easily be likened to the historic work carried out by information science on thesauri, multi-faceted language, or to computer science with ontologies. On the other hand, few researchers have really studied the genre or type of document that can be found on the Web. One can nevertheless quote the work of Crowston and Williams (2000), and Glover (2001) who are primarily interested in the notion of “genre” or type of document on the Web. The former studies the types of resources reproduced or emerging from the Web, such as FAQ or *home pages*. The latter presents an automated method which is able to recognize certain types of documents (personal homepages, calls for papers). More recently Kwasnik et al. (2001) studied how an information search could be improved by taking into account the type of Web documents.

Before attempting to study the genre and type of document available on the Web, we must reflect on the very nature of the Web document. The basic information unit retrieved by the majority of information retrieval systems available on the Web is the page. These units constitute Web hypertext network nodes and are basic components expressing a limited number of ideas (Balpe et al., 1996). They are self-reliant and stand on their own, but do not necessarily correspond to an entire document. Reading such pages is not always sufficient to understand and take in the document of which they are part, nor to index it correctly, i.e. to answer the following questions: what is this document about?, for which user, for what purpose? etc. It is difficult and even pointless to try and define a Web document in traditional terms, even though there are homogeneous sets of pages on the Web which can be easily identified. This is the case with Web sites which are coherent sets of pages (common objectives and themes), created and maintained by the same authority. As far as form is concerned, the pages of a site share the same graphic charter and sites always have a home page, an entry point for accessing the resources of the site. We have chosen first to characterize the Web sites and then the Web pages with the following three metadata elements: the type of authority responsible for the site, the type of site and the type of information contained on the page. The typology proposed is a personal approach that can evolve.

- Type of Authority: a better understanding of the informational content of the site and knowing who is responsible for its existence can be very useful. We identified four types of authority: *institutions*, *companies*, *associations* and *individuals*.
- Type of Site: This depends on the informational role that the site wishes to play. We have identified four distinct types.
    - The most common, the "shop-window" site (*home server*), contains mostly self-descriptive information, describing the authority responsible for the site. It is a type of "brochure" whose primary objective is presentation. The main topics of the site are: Who are we, our activities, products and partners, how to contact us, etc. However, deeper levels (several 'clicks' from the home page) of these sites can also contain documents which are not self-descriptive.
    - A *search site* provides access to Web resources. The most obvious examples are search engines and general directories. Specialized engines which list only a single type

of document such as Citeseer[1] (Lawrence et al., 1999) or a single medium type (such as image search engines) are also search sites.

- *Resource sites* perform an editorial function and unlike search sites, they organize and provide their own resources. They are often presented in the form of libraries or databases.
- *Web services* propose services related to life on the Web and the Internet, such as messaging systems, news forums, etc.

• Type of Information contained on the page: self-descriptive information, relating to the creator of the site, or non self-descriptive information.

4 Our Method: Propagation of Metadata values using the Co-citation Method

In 1998, Marchiori (1998) proposed a method to propagate (subject) classification metadata along the links. In this method, each page is described by subject metadata (keywords) weighted by a coefficient between 0 and 1 (1 if the metadata element exactly describes the page, 0 if it is inappropriate). His hypothesis is as follows: if a page P (described by a metadata element A weighted by coefficient ν) is cited on a page P', we can assume that P is used to clarify or support the ideas evoked on page P'. The metadata elements of P can therefore be back-propagated to P' with a weakening factor f ($0<f<1$). The metadata element A then describes the document P' with coefficient $\nu \times f$.
Marchiori’s hypothesis therefore stipulates that two pages connected by a hypertext link share common thematic metadata. This is no doubt true in a majority of the cases, but with certain limits that are familiar to all: publicity and navigation links. This hypothesis is not valid for other metadata such as the type of site (cf. paragraph 3). Servers that cite and that are cited are often of different types. A striking example: Web pages of search sites often cite Web pages coming from home servers or resource sites.

Like Marchiori, we believe that the Web graph is a vehicle of information. However, we would like to use a stronger relation than the

[1] http://citeseer.nj.nec.com/cs

simple association of “citing-cited”. The relation that interests us here is that of co-citation, that is to say the connection that exists between two pages cited together. If page P contains a hyperlink to pages P’ and P”, there is a reason, at least for the author of page P, to cite these two pages together. The existing association between the two pages P’ and P” is all the stronger if it is taken up by other authors and if pages P’ and P” are always cited together. Our hypothesis is that this association is rendered by identical values for one or several metadata. We have thus created a graph of co-citations (fig. 1) with which we have propagated metadata.

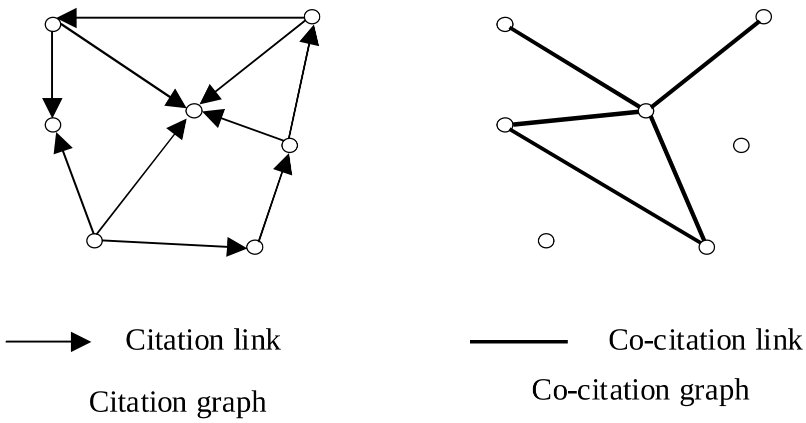

FIG. 1: Construction of a co-citation graph from a citation graph.

The method proposed involves two steps:

- Corpus clustering by the co-citation method to obtain a sub-corpus hierarchy which we assume to be homogeneous,
- The propagation of the metadata values in these sub-corpuses.

The method proposed allows propagation of the values of metadata for type of Site, type of Authority, type of Information contained on the page.

4.1 Corpus clustering by the co-citation method

The co-citation method has been used in bibliometrics since 1973

(Marshakova, 1973) (Small, 1973), and attempts to create relational maps of documents or authors from a set of scientific articles (or more precisely their bibliographic references) on a given subject in order to reflect both the sociological and thematic links in this field. This method is based on the hypothesis that two bibliographic references of any date which are frequently cited together have a thematic parity. The hypertext link itself can represent a citation, and several authors (Larson, 1996) (Pitkow and Pirolli, 1997) (Prime et al., 2002a) have been interested in transposing the document co-citation method to characterize the Web universe. They have brought out the theoretical and technical limits of analogy, but have shown the usefulness of the structuring process to bring together the subject content of the pages. Our method only takes into account inter-server links between citing and cited pages thereby hoping to reduce the number of navigation links.

The first phase of the method consists of determining how close the pages are to each other. To do this, we define a similarity index which aims to translate the following idea into a mathematical format: 2 pages P_i and P_j are close if their co-citation frequency (C_{ij}) is large with respect to their respective citation frequencies (C_i and C_j). There are several possible indices which, by convention, fall between 0 and 1: 1 when the pages are always cited together and 0 if they are never cited together. We have chosen the most common local index in scientometry called the equivalence index (Michelet, 1988).

$$E(i;j) = \frac{C_{ij}^2}{C_i C_j} \ .$$

We define $d_1(i;j)$ as the dissimilarity index associated with the equivalence index where $d_1(i;j)=1-E(i;j)$. The results are written in a co-citation matrix which represents the co-citation graph, a weighted graph where the nodes are the pages and the edges are the co-citation links weighted by d_1.

The second phase, the splitting of the co-citation graph to obtain homogenous groups, uses the methods of clustering from multi-dimensional analysis (Benzecri, 1973) (Hartigan, 1975). This is an agglomerative hierarchical clustering. Several agglomerative strategies are possible. The most conventional are the simple link (closest neighbour), the complete link (furthest neighbour), and the average link. This method is used to create a hierarchy of page clusters. The most similar documents are grouped in clusters at the lowest level, while at the top level all the documents are placed together. The hierarchy obtained can be viewed graphically by a dendrogram (fig. 2). In this study, we shall not determine specific cut-off level of dendrogram. Our goal is to use the whole dendrogram rather than a partition that would be obtained at a particular threshold level.We shall study every level in the dendrogram.

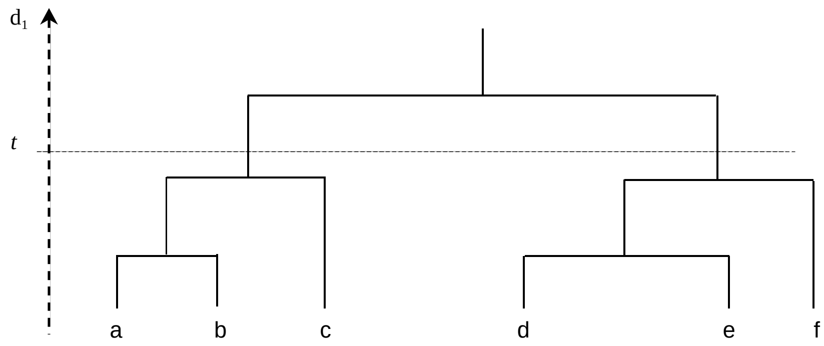

FIG.2: Example of a Dendrogram

4.2 Propagation in the Co-Citation Graph

For the starting group composed of *N* pages, we obtain a dendrogram. For each threshold *t* of the dendrogram corresponds a partition Π^t of clusters. The propagation method that we propose functions for a given threshold *t*. For each cluster of Π^t , composed of n pages, written $Cl = \{P_i / 1 \le i \le n\}$, the method consists of three stages which are described in the following sections.

4.2.1 Calculation of Page Centrality

Each cluster is a sub-graph of the co-citation graph. Our hypothesis is that the central element, that is to say the element that is the closest to all the others, is the most representative element of the cluster. The values of the metadata of this element are those that we want to propagate. That is why we classify the cluster pages according to their index of decreasing centrality. Centrality, a function introduced by Sabibussi (1966), is calculated in the following way :

$$\forall\ P_i \in\ Cl,\ \ Centrality(P_i) = \frac{n-1}{\sum_{j=1}^{n} d_2(P_i;P_j)}$$

where d_2 is the geodesic distance between two nodes P_1 and P_2 of a cluster, that is to say the sum of the arc valuations of the shortest route between P_1 and P_2. The centrality index varies between 0 and 1 and is equal to 1 when the element P_i is as central as possible, that is to say when it is adjacent to all the others. Several cluster pages can have the same centrality value.

4.2.2 Manual assignment of metadata values

In each cluster, we select the page (or one of the pages) that has the highest centrality value and then manually assign metadata to it. If one metadata element value cannot be determined, we select the next page classified in decreasing order of centrality and then we assign a value for this metadata element.

4.2.3 Propagation

The metadata values assigned previously are propagated to all the other pages of the cluster.

5. Experiment and Results

5.1 Creation of a Test Corpus

A hypertext corpus made up of qualified pages according to the metadata elements we have selected does not exist. That is why, in 2001, we created a test corpus containing 198 Web pages co-cited by 918 pages. We assigned metadata values to them manually according to the metadata elements related to the type of document as defined in section 4. The results of this manual assignment have been transcribed on a chart that, hereafter, we shall call the *chart of manual assignment*. Several times, we were not able to attribute metadata values. This was not due to an imprecise or incomplete typology, but rather to a lack of information available at the sites. This is often the case for resource or search sites where the authority is not always established, or for sites with an informational role that is not clearly ascertained. That is why our manual assignment chart contains undetermined values.

This corpus contains French language pages pertaining to astronomy. It was created thanks to the search engines Google[2] and Hotbot[3] using the query "astronomie" found only in French pages. We received 1541 different Web pages. To apply the co-citation method, the "father" pages of each of the 1541 pages - that is to say all of the pages that point to the latter pages - had to be found. 18714 father pages were found thanks to the *link* function provided by the search engines Google and Hotbot. Of the 1541 original pages, only 198 were co-cited.

The results of this manual assignment are presented in the chart below.

Type of authority		Type of site		Type of information	
Association	57	Home server	125	Self explanatory	104
Company	42	Research site	22	non self explanatory	78
Institution	39	Resource site	39	undetermined	16
Person	37	Web service	5	total	198
undetermined	23	undetermined	7		
total	198	total	198		

TABLE 1: Quantitative results of the manual assignment chart

5.2 Propagation

We tested our propagation method on the corpus. The 198 pages were grouped together in clusters using the co-citation method with the three possible strategies (simple link, average link, complete link).

The propagation of the selected metadata elements was realised for all three methods and for each of the thresholds. The results obtained for

[2] `http://www.google.com`
[3] `http://hotbot.lycos.com`

the average link strategy seemed to us to be the most significant, so we shall limit ourselves to these results in this article.

We shall examine the propagation results for each threshold *t*. The corpus contained N pages (in our experiment N=198) and so 3N metadata values. For each threshold *t*, we compared the metadata values obtained after applying the propagation method with those obtained using manual assignment (manual assignment chart). The propagation method split the 3N metadata values into 4 cells:

.

Metadata values	propagated	Not propagated
correct	a_c^p	a_c^{non-p}
incorrect	a_{inc}^p	a_{inc}^{non-p}

TABLE 2: Breakdown of the metadata values

- $a_c^p + a_{inc}^p$ is the number of propagated metadata values. a_c^p is the number of correct propagated metadata values, that is to say identical to those obtained by manual assignment. a_{inc}^p is the contrary.
- a_c^{non-p} is the number of metadata values assigned manually to propagate them to the other cluster pages. It must be noted that the number can be slightly higher than three times the number of clusters of Π^t since it is possible for a central element to carry an undetermined value.
- a_{inc}^{non-p} is the number of metadata values that were not assigned (either by propagation or by hand). This was the case for singleton pages. This number was foreseeable thanks to the dendrogram: it was equal to three times the number of singleton pages at threshold *t*.

It must be noted that $a_c^{non-p} + a_c^p + a_{inc}^p$ corresponds to the number of assigned metadata values.

3. **Presentation of the Results**

To interpret the results we defined three indices varying between

0 and 1.

- Propagation Quality

$$Q = \frac{a_c^p}{a_c^p + a_{inc}^p}$$

This is the ratio between the number of correctly propagated values and the total number of propagated values. This indicator measures the precision of the propagation. At the same time, it reflects the cohesion within the clusters.

.

- Performance

$$Perf = \frac{a_c^p + a_{inc}^p}{a_c^p + a_{inc}^p + a_c^{non-p}} = \frac{a_c^p + a_{inc}^p}{3N - a_{inc}^{non-p}}$$

This gives an indication of the number of metadata values propagated compared with the total number of values assigned by hand and by propagation.

- Ratio of processed pages (manually and by propagation)

$$T = \frac{3N - a_{inc}^{non-p}}{3N}.$$

$3 \times N$ is the number of metadata values in our corpus. At the lowest level in the dendrogram, there were many singletons, that is, pages which cannot be processed by this method. This index was used to know, for a given threshold, how many pages were affected by the process of manual assignment and propagation compared with the total number of pages in the corpus. It is to be noted that the value of the index T was foreseeable. It can be calculated by studying the dendrogram: a_{inc}^{non-p} is equal to three times the number of singletons.

In order to take into consideration the relationship between quality and performance, we charted a graph (figure 3) which presents the performance *Perf* as a function of the quality *Q* for the average link method. Note that the performance varied between 0.54 and 0.99 and the quality from 0.32 to 1 (value of 1 obtained for the last 4 thresholds).

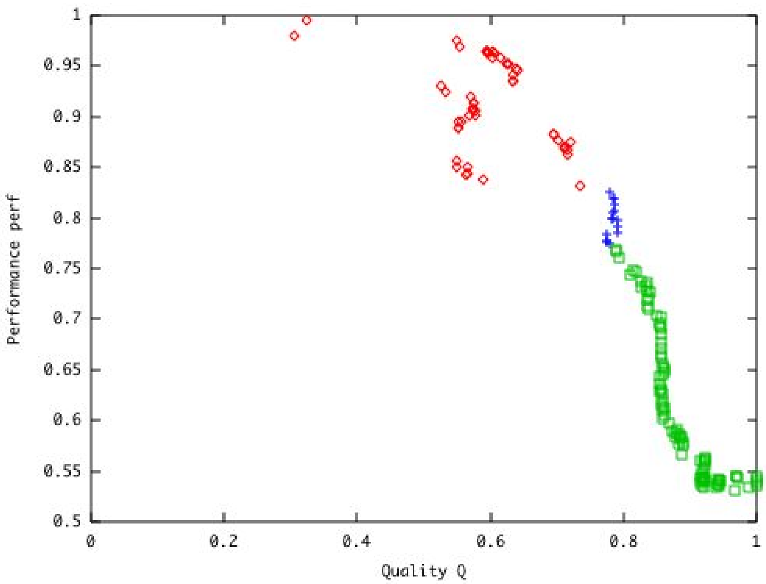

FIG.3: Diagram of performance as a function of quality

On the left of this graph, there is a group of dots with poor quality ($0.3 \leq Q \leq 0.77$) and good performance ($0.83 \leq Perf < 1$). These are the results obtained with the highest thresholds in the dendrogram. These thresholds had a low number of clusters, which is why their performance was high. On the other hand, the size of their clusters was quite large, their cohesion poor, the clustered pages did not have anything in common, which is what led to poor quality propagation. Note that in this zone, the quality did not depend on the performance. In fact, the clusters were not always stable, and the central element was not systematically representative of the majority of the pages. This is why the quality varied from one threshold to another.

Once the quality approaches 0.8, the graph shows a dependence between quality and performance. We note that the performance decreases as the quality increases. The clusters become smaller and smaller but more and more homogenous. This curve shows us that it is not possible to have perfect quality and perfect performance at the same time, which was predictable. There is however a very interesting region in which the thresholds have a

quality approaching 0.8 and a performance greater than 0.8.
The performance index that we selected was not sufficient to measure propagation. In fact, we can obtain good performance (low a_c^{non-p} compared to $a_c^p + a_{inc}^p$), but with a very high number of non-assigned values (a_{inc}^{non-p}) compared to the total number of metadata values of the corpus ($3 \times N$). The graph below shows the ratio of processed pages as a function of performance. Note that the number of processed pages varies between 0.11 and 1. The curve represented climbs very quickly in the performance zone [0.5;0.6] and then climbs regularly until it reaches 1. The zone that interests us the most with a quality and performance of about 0.8 corresponds to a remarkable ratio of processed pages of approximately 0.85 (figure 4).

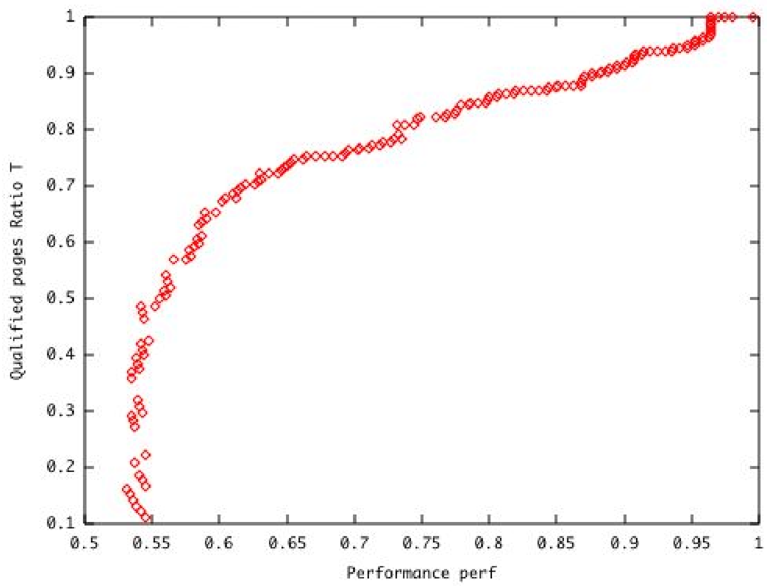

FIG.4: Ratio of processed pages as a Function of Performance

These two curves show that it is possible to obtain some thresholds for which the propagation quality is good (over 80%) and performance and ratio of processed pages that are acceptable (also over 80%). If these results can be confirmed with other experiments, the ratio of processed pages which is foreseeable (directly related to clusterisation), could become a cut-off criterium for clusterisation.

6. Conclusion and Limits

In this article, we have looked at the semi-automatic categorization of Web pages for three metadata elements related to the type (genre) of document: the site type, the organisation type and the information type. The method proposed used the co-citation graph. The results observed for quality, performance and ratio of processed pages are encouraging. In fact, we observed a number of thresholds for which we obtained values greater than or equal to 80% for these three indices. However, it appears to be impossible to categorize a majority of the pages of a corpus for these three metadata without introducing errors. It would undoubtedly be very interesting to test the same propagation method for other metadata such as the subject for example.

On the one hand we must acknowledge that the small size and the homogenous nature of our corpus is a limitation of our experiment. On the other hand, one of the limits of the method is without doubt the low proportion of pages co-cited on the Web. Many pages are not cited by pages hosted on other sites, so that they cannot be co-cited and classified. It is therefore necessary to devise a categorization method based on the propagation of metadata within Web sites. Note also that our similarity (equivalence) index is unrelated to the number of links on the citing pages. However, on the Web the number of links contained on each page varies significantly and it would be judicious to take this into account to calculate the closeness of pages. Currently, we are starting a larger experiment on a corpus containing 5 million pages corresponding to the French-language Web in December 2000 (collected by M. Géry and D. Vaufreydaz from the CLIPS laboratory `http://www-clips.imag.fr`). This experiment should allow us to become aware of any problem of scale which might be introduced and to clearly identify the percentage of pages co-cited and the number of pages that can thus be classified.

References

Balpe, J., Lelu, A., Papy, F., and Saleh, I. (1996). Techniques avancées pour l’hypertexte. éditions Hermès.

Benzecri, J.P. (1973). L'analyse de données, Tome 1 et 2, Dunod

Edition.
Björneborn, L. and Ingwersen, P. (2001). Perspectives of webometrics. Scientometrics, 50(1): 65–82.
Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In Proceedings of the Seventh International WWW Conference. IW3C2.
Crowston, K. and Williams, M. (2000). Reproduced and emergent genres of communication on the world wide web. The Information Society, 16(3): 201–215.
Ding, J., Gravano, L., and Shivakumar, N. (2000). Computing geographical scopes of web resources. In Proceedings of the 26th International Conference on Very Large Databases (VLDB 2000).
Dublin core metadata initiative (2003). at http: //dublincore.org, visited in February 2003.
Egghe, L. (2000). New informetric aspects of the internet : some reflections, many problems. Journal of information science, 26(5): 329–335.
Garfield, E. (1972). Citation analysis as a tool in journal evaluation. Science, (178): 471–479.
Glover, E., Flake, G., Lawrence, S., Birmingham, W., Kruger, A., Giles, L., and Pennock, D. (2001). Improving category specific web search by learning query modifications. In Symposium on Applications and Internet, SAINT 2001, San Diego, California.
Gravano, L. (2000). Characterizing web resources for improved search. In Position paper for the First NSF-DELOS Workshop on Information Seeking, Searching, and Querying in Digital Libraries.
Hartigan, J. (1975) Clustering Algorithms. John Wiley and Sons, New York.
Ingwersen, P. (1998). The calculation of web impact factors. Journal of Documentation, 54(2): 236–243.
Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. (1999). Trawling the web for emerging cyber-communities. In Proceedings of the Eighth World Wide Web Conference.
Kwasnik, B., Crowston, K., Nilan, M., and Roussinov, D. (2001). Identifying document genre to improve web search effectiveness. The Bulletin of the American Society for Information Science and Technology, 27(2).
Larson, R. (1996). Bibliometrics of the world wide web : An exploratory analysis of the intellectual structure of the cyberspace. In Proceedings of the Annual Meeting of the American Society of Information Science, Baltimore.

Lawrence, S., Bollacker, K., and Giles, C. (1999). Indexing and retrieval of scientific literature. In Proceedings of the Eighth International Conference on Information and Knowledge Management, CIKM 99, pages 139–146.

Marchiori, M. (1998). The limits of web metadata and beyond. In Proceedings of the Seventh International WWW Conference. IW3C2.

Marshakova, I. V. (1973). Document coupling system based on references taken from science citation index. Russia, Nauchno - Teknicheskaya Informatsiya, 2(6,3).

Michelet, B. (1988) L'analyse des associations. Thèse de doctorat, Université de Paris VII, UFR de Chimie, Paris, 26 Octobre 1988. Spécialité: Information Scientifique et Technique.

Olsina, L., Lafuente, G., and Rossi, G. (2001). Specifying quality characteristics and attributes for websites. Lecture Notes in Computer Science, 2016

Pitkow, J. and Pirolli, P. (1997). Life, death and lawfulness on the electronic frontier. In Proceedings of the ACM SIGCHI Conference on Human Factors in Computing System, CHI'97, pages 118–125.

Prime, C., Bassecoulard, E., and Zitt, M. (2002a). Co-citations and co-sitations: a cautionary view on an analogy. Scientometrics, 54(2): 291–308.

Prime, C., Beigbeder, M., and Lafouge, T. (2002b). Clusterisation du web en vue d'extraction de corpus homogènes. In Actes du 20ème congrès INFORSID, pages 229–242, Nantes.

Small, H. (1973). Co-citation in the scientific literature. Journal of the American Society for Information Science, 24: 265–269